

# PATTERNS OF LINKAGE DISEQUILIBRIUM IN THE HUMAN GENOME

Kristin G. Ardlie<sup>\*</sup>, Leonid Kruglyak<sup>‡</sup> and Mark Seielstad<sup>§</sup>

Particular alleles at neighbouring loci tend to be co-inherited. For tightly linked loci, this might lead to associations between alleles in the population — a property known as linkage disequilibrium (LD). LD has recently become the focus of intense study in the hope that it might facilitate the mapping of complex disease loci through whole-genome association studies. This approach depends crucially on the patterns of LD in the human genome. In this review, we draw on empirical studies in humans and *Drosophila*, as well as simulation studies, to assess the current state of knowledge about patterns of LD, and consider the implications for the use of LD as a mapping tool.

The past successes of disease-gene mapping owe much to the fact that the first genetic diseases to be studied in humans were fairly simple, monogenic, highly penetrant disorders that obey the rules of Mendelian inheritance. Most were identified by linkage analyses, in which data are collected from affected families, and regions of the genome are identified that co-segregate with the disease in many independent families or over many generations of an extended pedigree. The disease locus, if there is indeed just one relevant locus in the entire collection of families, will lie in the region of the genome that is shared by all affected members of a pedigree. Generally, genes can be localized only within a very large interval by this approach, because the size of the co-segregating piece of DNA can only be delimited by the observation of a crossover between a marker and the disease locus itself. Unless one had numerous families, or very large, multigeneration pedigrees, the number of observed crossovers is small, with the resulting gene being mapped to a consequently large interval<sup>1,2</sup>. This was a particularly important limitation in the days when very little of the genome had been sequenced.

Other methods were therefore sought for narrowing the interval in which a disease gene might lie, and one of these was by the analysis of linkage disequilibrium (LD).

If most affected individuals in a population share the same mutant allele at a causative locus, it is possible to narrow the genetic interval around the disease locus by detecting disequilibrium between nearby markers and the disease locus<sup>3–5</sup> (FIG. 1). This approach makes use of the many opportunities for crossovers between markers and the disease locus during the large number of generations since the first appearance of the mutation. This interest in LD led to the first studies of its magnitude and behaviour in humans<sup>5–9</sup>. However, the early applications of LD mapping were limited to rare diseases in a few favourable populations.

More recently, human geneticists have turned their efforts to more common diseases that affect large swathes of the population and have a complex genetic basis. Here, family linkage studies have been much less successful. As a result, there has been a resurgence of interest in LD, owing largely to the belief that association studies offer substantially greater power for mapping common disease genes than do traditional linkage studies, and that LD can offer a shortcut to genome-wide association studies. This interest has been bolstered by the progress that has been made in the sequencing of the human genome and the establishment of large collections of single-nucleotide polymorphisms (SNPs), such as those identified by the **SNP Consortium**<sup>10</sup>.

<sup>\*</sup>Genomics Collaborative, 99 Erie Street, Cambridge, Massachusetts 02139, USA.  
<sup>‡</sup>Howard Hughes Medical Institute and Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, D4-100, Seattle, Washington 98109, USA.  
<sup>§</sup>Program for Population Genetics, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115, USA.  
Correspondence to K.G.A., L.K. or M.S. e-mails: [kardlie@genomicsinc.com](mailto:kardlie@genomicsinc.com); [leonid@fhcr.org](mailto:leonid@fhcr.org); [mark@ppg.harvard.edu](mailto:mark@ppg.harvard.edu)  
DOI: 10.1038/nrg777

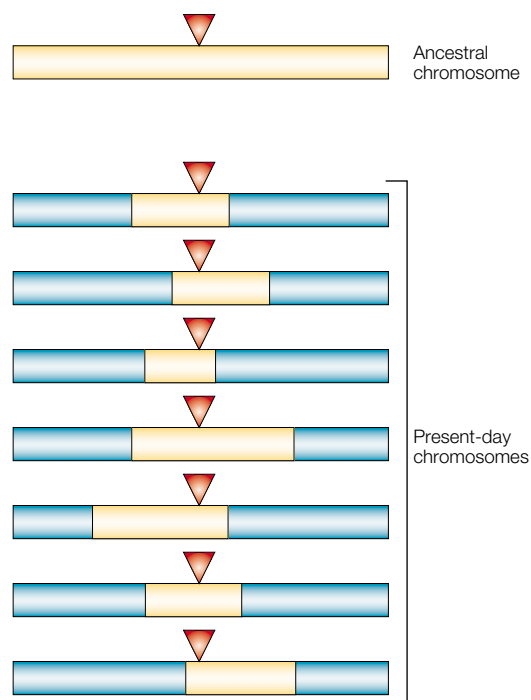


Figure 1 | **Linkage disequilibrium around an ancestral mutation.** The mutation is indicated by a red triangle. Chromosomal stretches derived from the common ancestor of all mutant chromosomes are shown in yellow, and new stretches introduced by recombination are shown in blue. Markers that are physically close (that is, in the yellow regions of present-day chromosomes) tend to remain associated with the ancestral mutation even as recombination limits the extent of the region of association over time.

An analysis by Risch and Merikangas<sup>11</sup> suggested that, in a linkage study, the number of pedigrees required to map the genes of minor effect that probably underlie susceptibility to common diseases would be prohibitively large. In these circumstances, for common disease alleles, they advocated the use of population-based TESTS OF ASSOCIATION (FIG. 2). In their conception, several unrelated markers in every known gene would be tested for an association with the disease. This in itself would entail the genotyping of a vast number of markers, and there would be no guarantee that the causative variant would be included. The number of markers to be tested could range from the ~50,000 nonsynonymous coding SNPs (cSNPs) to the ~7 million SNPs with both alleles above 5% frequency<sup>12</sup>.

An immediate question is whether a susceptibility locus could be implicated by detecting an indirect association, through LD, between a nearby marker and the disease. If disequilibrium were extensive, the number of markers used in a genome-wide test of association could be reduced, without an unacceptable probability of missing the association. Even variants that are not present in the screen would be assayed indirectly through LD with nearby markers. This approach is known as LD mapping. If this strategy is to be applied successfully, we need to understand the behaviour of LD and, in particular, gain some insight into how far usable

levels of disequilibrium extend in the human genome, and how much this varies from one region or population to another. Answers to these questions will determine whether LD mapping of susceptibility genes can be carried out with a feasible number of markers, although feasibility is an ever-moving target as SNP genotyping technologies improve<sup>13</sup>.

In this review, we discuss our current understanding of the levels and patterns of LD in the human genome. Initially, the potential for the application of LD in humans was assessed using simulations, but over the past two years, numerous publications have reported empirical studies that have examined LD over defined regions of the genome. We begin by providing some background information that covers the theory of LD, as well as some of the early empirical data from *Drosophila*, which lays the foundation for considering the more recent human studies and their implications for efforts to understand the genetic basis of common diseases.

### The nature of linkage disequilibrium

LD is the non-random association of alleles at adjacent loci. When a particular allele at one locus is found together on the same chromosome with a specific allele at a second locus — more often than expected if the loci were segregating independently in a population — the loci are in disequilibrium. This concept of LD is formalized by one of the earliest measures of disequilibrium to be proposed (symbolized by  $D$  (REF. 14)).  $D$ , in common with most other measures of LD (see below), quantifies disequilibrium as the difference between the observed frequency of a two-locus HAPLOTYPE and the frequency it would be expected to show if the alleles were segregating at random. Adopting the standard notation for two adjacent loci — **A** and **B**, with two alleles ( $A, a$  and  $B, b$ ) at each locus — the observed frequency of the haplotype that consists of alleles  $A$  and  $B$  is represented by  $P_{AB}$ . Assuming the independent assortment of alleles at the two loci, the expected haplotype frequency is calculated as the product of the allele frequency of each of the two alleles, or  $P_A \times P_B$ , where  $P_A$  is the frequency of allele  $A$  at the first locus and  $P_B$  is the frequency of allele  $B$  at the second locus. So, one of the simplest measures of disequilibrium is

$$D = P_{AB} - P_A \times P_B \quad (1)$$

An easy way to imagine how LD arises is to consider how polymorphisms are generated by mutation. LD is created when a new mutation occurs on a chromosome that carries a particular allele at a nearby locus, and is gradually eroded by recombination (FIG. 3). Recurrent mutations (for example, at either of the two loci in FIG. 3) can also lessen the association between alleles at adjacent loci. However, for SNPs, if not for microsatellites, recurrent mutations are generally very rare, and there is no evidence to indicate that mutation contributes significantly to the erosion of LD between SNPs.

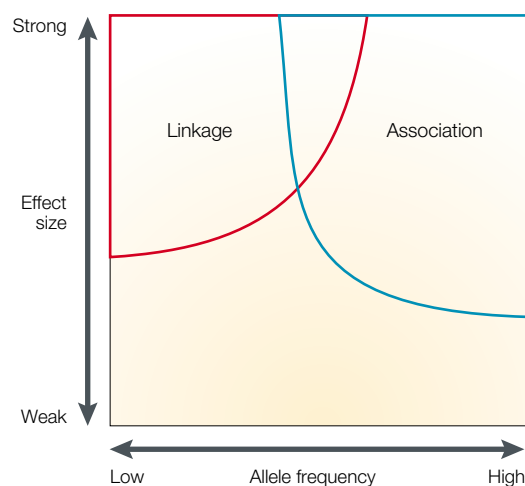
**TEST OF ASSOCIATION**  
A study that searches for a population association between a phenotype and a particular allele. In one of the simplest such tests — the case-control study — an excess (or deficiency) of a particular allele in a sample of individuals with the phenotype is looked for, as compared with a sample without the phenotype drawn from the same population.

**HAPLOTYPE**  
The combination of alleles found at neighbouring loci on a single chromosome or haploid DNA molecule.

The importance of recombination in shaping patterns of LD is acknowledged by the moniker of ‘linkage’. The extent of LD in populations is expected to decrease with both time ( $t$ ) and recombinational distance ( $r$ , or the recombination fraction) between markers. Theoretically, LD decays with time and distance according to the following formula, where  $D_0$  is the extent of disequilibrium at some starting point and  $D_t$  is the extent of disequilibrium  $t$  generations later:

$$D_t = (1 - r)^t D_0. \quad (2)$$

In fact, this deterministic equation poorly represents the behaviour of LD over short distances, where stochastic factors predominate. Consequently, although a trend towards decreasing disequilibrium with increasing distance between markers has generally been observed in empirical data, closely ‘linked’ markers are not always in LD<sup>15–18</sup>. By contrast, in other instances, LD has been reported between quite distant markers<sup>19–22</sup>. This variability underscores the conclusion that the factors governing LD among any specific collection of loci are numerous, complex and sometimes only dimly understood. A range of demographic, molecular and evolutionary forces (BOX 1) have a significant effect on the patterns of LD. Although empirical studies of these factors are in their infancy, simulations are beginning to illustrate their effects on LD, as we discuss in greater detail later.



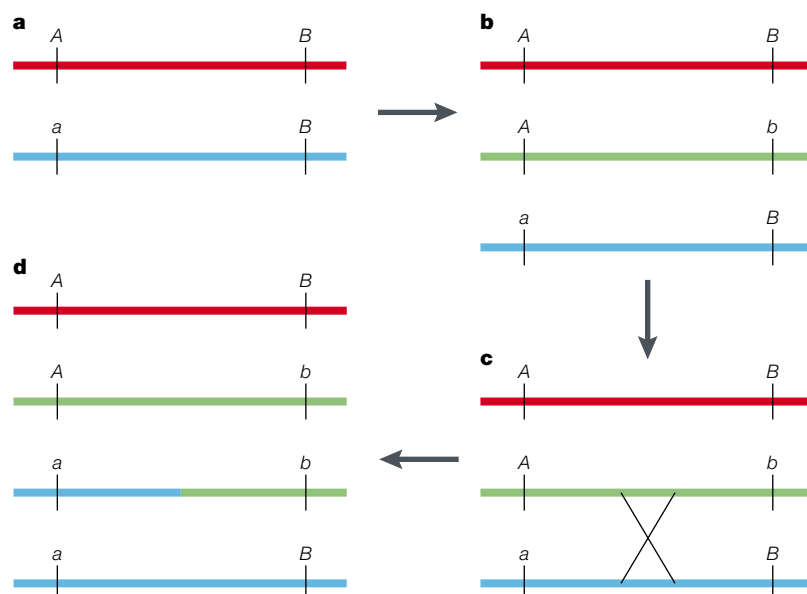
**Figure 2 | Optimal mapping strategies for different types of loci.** This schematic shows the relative efficiency of the two primary trait-mapping approaches in current use. Generally, tests of association are more powerful than linkage studies when the predisposing variant is frequent. The allele frequency is rarely known in advance, which has fuelled much speculation on the likely frequency of allelic variants for common disease<sup>70,72</sup>. Although there is some support for “common variants that underlie common disease”, there is still much debate and few empirical examples to resolve it, and no single answer is likely. The precise magnitude of the detectable effect size in any particular study will depend importantly on the sample size. With increasing sample sizes, alleles of increasingly minor effect can be identified. Note that for alleles of weak effect, all current approaches are inadequate.

### Measures of linkage disequilibrium

Although the measure  $D$  captures the intuitive concept of LD, its numerical value is of little use for measuring the strength of and comparing levels of LD. This is due to the dependence of  $D$  on allele frequencies. As a result, several alternative measures based on  $D$  have been devised (reviewed in Devlin and Risch<sup>23</sup>). Comparing different reports on the extent of LD is complicated by the fact that several measures are in common use, and although all are based on Lewontin’s  $D$ , they have very different properties and measure different things. The two most common measures are the absolute value of  $D'$ , and  $r^2$ .

The absolute value of  $D'$  is determined by dividing  $D$  by its maximum possible value, given the allele frequencies at the two loci. This has the useful property that  $D' = 1$  if, and only if, two SNPs have not been separated by recombination (or recurrent mutation or gene conversion) during the history of the sample. In this case, at most three out of the four possible two-locus haplotypes are observed in the sample (FIG. 3). The case of  $D' = 1$  is known as complete LD. Values of  $D' < 1$  indicate that the complete ancestral LD has been disrupted. However, the relative magnitude of values of  $D' < 1$  has no clear interpretation. In addition, estimates of  $D'$  are strongly inflated in small samples, even for SNPs with common alleles, but especially for SNPs with rare alleles. So, high values can be obtained even when markers are in fact in linkage equilibrium. Because the magnitude of  $D'$  depends strongly on sample size, samples are difficult to compare. Therefore, statistically significant values of  $D'$  that are near one provide a useful indication of minimal historical recombination, but intermediate values should not be used for comparisons of the strength of LD between studies, or to measure the extent of LD.

The measure  $r^2$  (sometimes also denoted by  $\Delta^2$ ) is in some ways complementary to  $D'$ , and has recently emerged as the measure of choice for quantifying and comparing LD in the context of mapping<sup>24,25</sup>. It is the correlation of alleles at the two sites, and is formed by dividing  $D^2$  by the product of the four allele frequencies at the two loci.  $r^2 = 1$  if, and only if, the markers have not been separated by recombination and have the same allele frequency. In this case, exactly two out of the four possible two-locus haplotypes are observed in the sample (such as  $A,B$  and  $a,b$  in FIG. 3). The case of  $r^2 = 1$  is known as perfect LD. In this case, observations at one marker provide complete information about the other marker, making the two redundant. Intermediate values of  $r^2$  are easily interpretable. Consider two loci: one locus is functionally associated with disease and the other is a nearby marker in LD with the susceptibility locus. To have the same power to detect the association between the disease and the marker locus, the sample size must be increased by roughly  $1/r^2$  when compared with the sample size for detecting association with the susceptibility locus itself<sup>24,26</sup>. Put more simply, the value of  $r^2$  is related to the amount of information provided by one locus about the other. Notably, this property correctly takes into account differences in allele frequencies



**Figure 3 | The erosion of linkage disequilibrium by recombination.** **a** | At the outset, there is a polymorphic locus with alleles *A* and *a*. **b** | When a mutation occurs at a nearby locus, changing an allele *B* to *b*, this occurs on a single chromosome bearing either allele *A* or *a* at the first locus (*A* in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The *b* allele will always be found on a chromosome with the *A* allele at the adjacent locus. **c** | The association between alleles at the two loci will gradually be disrupted by recombination between the loci. **d** | This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (*a*, *b*) increases in frequency.

at the two loci. However, it also means that two markers that are immediately adjacent might show different  $r^2$ -values with a third marker, and that a low pairwise  $r^2$ -value is not necessarily indicative of high ancestral recombination in the region.  $r^2$  also shows much less inflation in small samples than does  $D'$  (REF. 25).

The interpretation of  $r^2$  in terms of the power to detect an association leads to the concept of useful LD<sup>26</sup>. Sample size is usually limiting in association studies (because of the cost and effort that are involved in patient recruitment, phenotyping and genotyping), and large increases in sample size to compensate for weak LD between a marker and the susceptibility locus are impractical. Values of  $r^2 > 1/3$  limit the required increase in sample size to no more than threefold, and should probably be taken to be the minimum useful values. Much higher  $D'$ -values are generally needed to indicate similarly useful levels of LD, because of the tendency of  $D'$  to overestimate the magnitude of LD. In particular, the 'half-length' of  $D'$  (the distance at which it falls to 0.5)<sup>21</sup> greatly overstates the range over which LD is useful for mapping<sup>25</sup>.

Additional confusion arises from descriptions of LD in terms of  $p$ -values in a test of significant departure from linkage equilibrium between loci. Because  $p$ -values depend strongly on sample size, they cannot be used to compare LD between studies with different sample sizes. Furthermore, even very low levels of LD can be statistically significant in a sufficiently large sample. For example, an  $r^2$ -value of 0.01 can be statistically

significant in a sample of 1,000 chromosomes. Because weak deviations from linkage equilibrium might extend over considerable distances,  $p$ -values can create a misleading impression that LD extends over great distances, when the actual level of LD present at such distances is typically far below that which is useful for mapping.

Another approach for quantifying LD is through the population recombination parameter  $4N_e r$  (alternatively denoted by  $\rho$ ,  $4N_e c$  or  $C$ ), where  $r$  or  $c$  is the recombination rate across the region of interest and  $N_e$  is the EFFECTIVE POPULATION SIZE<sup>24</sup>. This approach avoids reliance on pairwise measures of LD, which differ from marker to marker, and facilitates comparisons between regions. In practice, estimation of  $4N_e r$  from genotyping data is computationally challenging — the theory of optimal estimation is not fully worked out — and estimators rely on assumptions about demography and selective neutrality<sup>24,27–33</sup>. The theoretical and intuitive appeal of  $4N_e r$  as a measure of the extent of historical recombination in a region is certain to keep this an active area of research, and the use of this measure should become increasingly common as methods to compute it improve.

In summary, current research strongly favours the use of  $r^2$  as a pairwise measure of LD in the context of association studies. As a rough rule of thumb,  $r^2$ -values above 1/3 might indicate sufficiently strong LD to be useful for mapping. Statistically significant values of  $D'$  that are near one can indicate regions of low historical recombination, but intermediate values of  $D'$  should be avoided as measures of LD. The population recombination parameter  $4N_e r$  shows considerable promise for quantifying the strength of LD in a region, pending further theoretical and computational development.

### Linkage disequilibrium in *Drosophila*

Much of our understanding of how LD is shaped in natural populations initially came from research on *Drosophila* species. Historically, the most detailed analyses of LD have been made in *Drosophila melanogaster*, in which allelic combinations (haplotypes) can readily be determined for individual chromosomes that have been extracted from wild populations through inbred lines. Most studies have focused on in-depth comparisons of single gene loci and/or single populations, and the principal finding is one of regional variation in LD among loci. Detailed comparisons among gene regions have revealed large differences in the intensity of LD in different regions<sup>34</sup>, with variation in both recombination and the number of gene conversion events that are implicated as the cause<sup>35,36</sup>. In-depth studies of how several forces (for example, mutation, recombination and selection) act to increase or decrease LD in a given region indicate that the balance of these forces should result in strongest disequilibrium around alleles at frequencies of ~10% (REF. 37). However, even adjacent regions can experience quite different evolutionary histories. A recent chromosome-wide study of the fourth chromosome, previously believed to be non-recombining and invariable, found polymorphic regions interspersed with regions of little to no variation<sup>38</sup>. Recombination

**EFFECTIVE POPULATION SIZE ( $N_e$ ).** An abstraction, equivalent to the breeding population size of an ideal, randomly mating population, which maintains the same level of variation as observed in the actual population.  $N_e$  often bears little relation to the actual number of living or reproducing individuals (the census population size), and is generally much smaller.

was shown to occur on the chromosome, and although only at a very low rate consistent with previous findings, it has been sufficient to affect the structure of genetic variation on the chromosome, allowing different regions to have different evolutionary histories.

Recombination rates per physical length are well known to show marked regional variation, and much research on LD in *Drosophila* has used this fact to focus on understanding the effects of selection and other forces on the degree of LD. Over the past decade,

numerous surveys of DNA sequence variation in natural populations of several *Drosophila* species have established that polymorphism levels are positively correlated with the regional rate of crossing over, and are not generally explained by variation in mutation rates<sup>38–40</sup>. This correlation has been proposed to result from the hitchhiking that is associated with fixation of advantageous mutants: in a region of low recombination, if directional selection drives an advantageous mutation through a population to fixation, much of the

#### Box 1 | Factors that influence linkage disequilibrium

Mutation and recombination might have the most evident impact on linkage disequilibrium (LD), but there are additional contributors to the extent and distribution of disequilibrium. Most of these involve demographic aspects of a population, and tend to sever the relationship between LD strength and the physical distance between loci.

- **Genetic drift.** This phenomenon describes the change in gene and haplotype frequency in a population every generation owing to the random sampling of gametes that occurs during the production of a finite number of offspring. Frequency changes are accentuated in small populations. In general, the increased drift of small, stable (not growing) populations tends to increase LD, as haplotypes are lost from the population. Such populations might be suitable for disease-gene mapping, with the idea that genetic drift will accentuate disease and marker allele frequency differences between cases and controls<sup>75</sup>. However, the applicability of this phenomenon to gene mapping has not been well characterized.
- **Population growth.** Rapid population growth decreases LD by reducing genetic drift.
- **Admixture or migration.** LD can be created by ADMIXTURE, or by migration (gene flow), between populations. Initially, LD is proportional to the allele frequency differences between the populations, and is unrelated to the distance between markers. In subsequent generations, the ‘spurious’ LD between unlinked markers quickly dissipates, while LD between nearby markers is more slowly broken down by recombination. In theory, this would allow the mapping of disease genes in hybrid populations without using many genetic markers<sup>76</sup>. Several admixed populations, such as African Americans and Hispanic Americans, have been characterized with this application in mind<sup>77–79</sup>, but the success of this approach will depend heavily on the time since admixture occurred, the frequency differences of the disease of interest in the parental populations and the allele frequency differences. So, the diseases and circumstances for which this mapping approach will be feasible might turn out to be quite rare and exceptional.
- **Population structure.** Various aspects of population structure are thought to influence LD. Population subdivision is likely to have been an important factor in establishing the patterns of LD in humans, but most of our limited knowledge comes from the study of model organisms. An interesting recent study of *Arabidopsis* indicated that extreme inbreeding can produce high levels of LD without a substantial reduction in levels of variation<sup>80</sup>. This neglected area would benefit from intensified study in humans.
- **Natural selection.** There are two primary routes by which selection can affect the extent of disequilibrium. The first is a hitchhiking effect, in which an entire haplotype that flanks a favoured variant can be rapidly swept to high frequency or even fixation<sup>38,41,42</sup>. Although the effect is generally milder, selection against deleterious variants can also inflate LD, as the deleterious haplotypes are swept from the population<sup>43</sup>. The second way in which selection can affect LD is through epistatic selection for combinations of alleles at two or more loci on the same chromosome<sup>81</sup>. This form of selection leads to the association of particular alleles at different loci. Although this has provided a major motivation for historical studies of LD in *Drosophila* genetics, as a means of detecting the action of (epistatic) natural selection<sup>14</sup>, it has not yet been shown to alter LD in humans.
- **Variable recombination rates.** Recombination rates are known to vary by more than an order of magnitude across the genome. Because breakdown of LD is primarily driven by recombination, the extent of LD is expected to vary in inverse relation to the local recombination rate. It is even possible that recombination is largely confined to highly localized recombination hot spots, with little recombination elsewhere. According to this view, LD will be strong across the non-recombining regions and break down at hot spots. Although there are intriguing indications that this reflects the situation for some regions<sup>66</sup>, the generality of the hot-spot phenomenon, the strength of recombination in and outside hot spots, and the length distributions of these regions remain to be determined.
- **Variable mutation rates.** Some single-nucleotide polymorphisms, such as those at CpG dinucleotides, might have high mutation rates and therefore show little or no LD with nearby markers, even in the absence of historical recombination.
- **Gene conversion.** In a gene conversion event, a short stretch of one copy of a chromosome is transferred to the other copy during meiosis. The effect is equivalent to two very closely spaced recombination events, and can break down LD in a manner similar to recombination or recurrent mutation. It has recently been shown that rates of gene conversion in humans are high and are important in LD between very tightly linked markers<sup>18,28,74</sup>.

#### ADMIXTURE

The introduction of mates (more properly, their genes) from one previously distinct population (the gene frequencies of which might differ) into another.



CLINE

The exhibition of regular and directional variation in phenotype, or genotype, across a geographical region — for example, steadily increasing pigmentation from temperate to tropical latitudes.

PANMIXIS

The process in a population by which mates are chosen at random with equal probability regardless of geographical location or genotype.

NUCLEOTIDE DIVERSITY

A measure of DNA sequence variation that is influenced both by the number of variable sites and their population allele frequencies in a given stretch of DNA.

MINOR ALLELE

The less frequent of two alleles at a locus.

BOTTLENECK

A period when the size of a population is reduced, generally by a large factor.

variation at linked sites will be eliminated during the process<sup>41</sup>. Selection on a region will therefore also increase the strength of LD observed: that is, significant allelic associations over large genetic distances might result from the action of natural selection. For example, strong geographical CLINAL variation in many enzyme loci around the phosphogluconate mutase (*Pgm*) locus is likely to be explained by clinal selection at *Pgm* and pervasive low levels of recombination in the region, so that the other loci are forced to hitchhike along with it<sup>42</sup>. Selection against deleterious mutations can also reduce nucleotide variation at linked sites<sup>43</sup>. A recent analysis of multiple loci in *D. melanogaster* and *Drosophila simulans* showed that both species have greater within-locus LD than expected theoretically<sup>44</sup>. This could be due to a departure from the demographic assumption of a PANMIXIC equilibrium in *Drosophila* and/or the action of natural selection on many of the loci.

Recombination rates in humans vary similarly by more than an order of magnitude throughout the genome<sup>45</sup>, as do the levels of heterozygosity for SNPs. Like the earlier findings in *Drosophila*, recent analyses indicate that the variation in levels of nucleotide polymorphism in humans might be explained by regional differences in the rate of recombination as well as by the action of natural selection<sup>46</sup>. To summarize, studies in *Drosophila* have illustrated that several factors, in addition to physical and genetic distance, can influence LD, and that region-to-region variability is the hallmark of LD, which presages the situation we are now seeing in humans.

**The importance of demographic history**

Because empirical data on the extent of LD between SNPs in humans were few when genome-wide association studies were first proposed, an initial attempt to estimate the necessary marker density in different populations used simulations<sup>26</sup>. This study first considered a simplified model of the human population, in which an ancestral population with an effective population size ( $N_e$ ) of 10,000 expanded exponentially, starting 5,000 generations ago, to a current size of 5 billion. Interestingly, despite the expansion, the NUCLEOTIDE DIVERSITY in the present-day population under this model is only 20% higher than in a population with a constant effective population size of 10,000. This is because 5,000 generations is too short a time to meaningfully alter genetic diversity in a population with a starting effective size of 10,000. If a useful level of LD is defined as one that allows a disease association to be detected through LD in a sample that is less than 2–3 times larger than that needed to detect the association directly (see above), then, under this model, a useful level of LD on average extends for 3–5 kb in either direction from a disease-causing variant. The study also showed that reducing the ancestral  $N_e$  to 1,000 while delaying the start of the expansion until 1,000 generations ago would increase the average extent of useful levels of LD by an order of magnitude to 30 kb in either direction from a variant. Such a model could reasonably represent a continental

population, such as that of Europe, but would have severalfold lower nucleotide diversity than experimentally observed. The study also confirmed that levels of LD show marked variability around the average, in accord with previous theoretical and empirical work.

The demographic models that underlie these results are highly simplified, and, as noted by Pritchard and Przeworski<sup>24</sup>, should produce a profound skew in the frequency spectrum of MINOR ALLELES, which has not been seen in actual data<sup>47</sup>. Indeed, in some instances, empirical data are more consistent with a constant population model<sup>24</sup> or a model of moderate growth in  $N_e$  within a subdivided population<sup>48</sup>. Pritchard and Przeworski<sup>24</sup> carried out simulations for several demographic models chosen to match the experimentally observed levels of nucleotide diversity in humans. They showed that population growth tends to decrease the extent of LD, especially for longer periods of growth. By contrast, population subdivision tends to increase the extent of LD, especially when a sample contains individuals from several strongly differentiated subpopulations.

Clearly, the true demographic history of the human population is very complex, with populations in different parts of the globe experiencing varying degrees of isolation, migration, admixture, expansion and BOTTLENECKS after the expansion of the ancestral African population. Unfortunately, the specifics are largely lost in the sands of time, and it is difficult to justify any model as a detailed reflection of reality. Instead, models can best serve to improve intuition about the importance and direction of effect of changes in different demographic parameters, and to establish what constraints are placed on ranges of parameters by empirical observations. In particular, the consistent observation of nucleotide diversity in the range of  $5\text{--}10 \times 10^{-4}$  in different populations<sup>22,47</sup> must be taken into account when considering demographic models applied to LD.

We carried out simulations for several models of demographic history, including bottlenecks and expansions (BOX 2). What clearly emerges is that the extent of LD is closely and inversely correlated with nucleotide diversity, and that any model with nucleotide diversity in the empirically observed range has a limited extent of LD, with useful levels of LD limited to ~10 kb in either direction from a variant (BOX 2d).

Models in which LD extends over much greater distances involve narrow bottlenecks that result in levels of nucleotide diversity that are much lower than those seen in actual human data. This observation complicates the interpretation of empirical reports of high levels of LD that extend for tens of kilobases or more, on average, in populations with typical levels of nucleotide diversity. If these reports are correct, simple demographic models cannot explain both LD and nucleotide diversity. The limits of such models have also been pointed out by others<sup>24,28</sup>. More complex

demographic models, perhaps involving differentiation of subpopulations and admixture among them, and/or heterogeneity in molecular factors such as recombination rate, might be necessary to fully explain the observations.

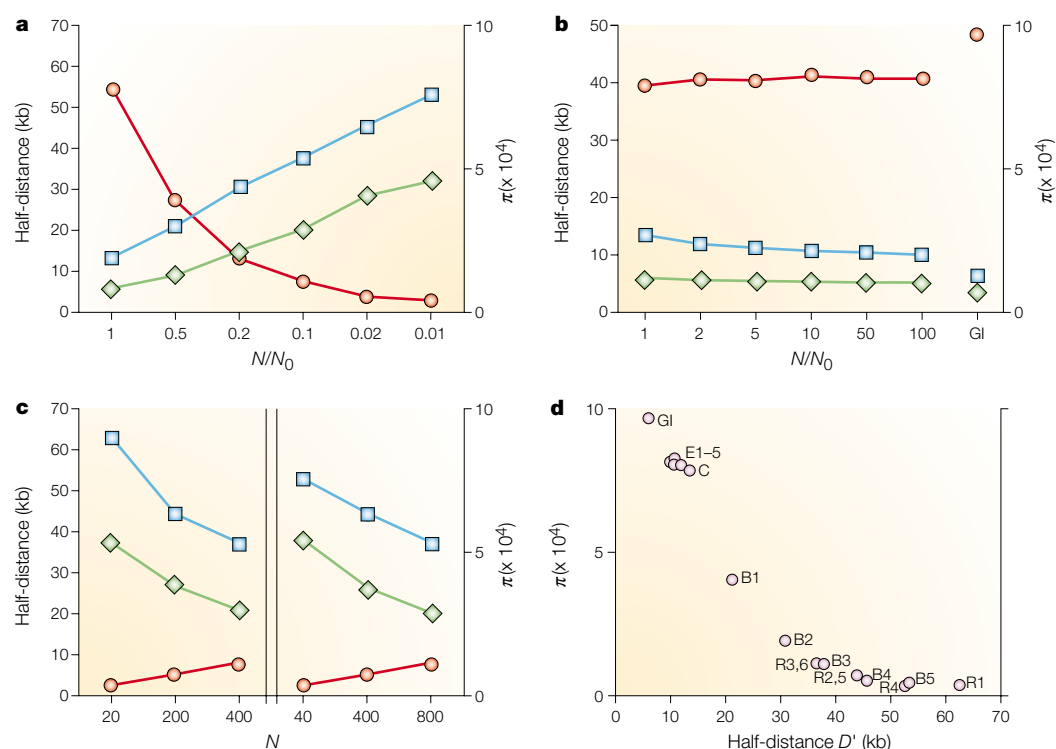
### Empirical human studies

In the past year and a half, the number of empirical studies that characterize the extent and range of LD in human populations has grown immensely. Most earlier studies comprised data sets of microsatellites<sup>49–51</sup>, but

#### Box 2 | Simulating the effects of demography on LD and nucleotide diversity

Demographic events, such as population bottlenecks and expansions, have effects on both the extent of linkage disequilibrium (LD) and the amount of sequence variation. Here, we illustrate the effects of several simple demographic models through simulations. For each model, we simulated the genealogy of 50 individuals drawn from the present-day population and computed two measures of LD for all pairs of sites with allele frequency >20%, as well as nucleotide diversity.

- **LD and nucleotide diversity for demographic models with a bottleneck.** (Panel a.) Distances where LD falls to half the maximum value for the metrics  $r^2$  (green diamonds) and  $D'$  (blue squares) are plotted against the population contraction factor during the bottleneck. The models have a constant effective population size of  $N_0 = 10,000$  up to 2,000 generations ago, at which time the population is reduced to a size  $N$  for 1,000 generations and then increases back to  $N_0$ , remaining at this size until the present. Nucleotide diversity ( $\pi$ , red circles) is also plotted for the different models. Note that the extent of LD increases with increasing severity of the bottleneck, whereas nucleotide diversity decreases.
- **LD and nucleotide diversity for demographic models with population expansion.** (Panel b.) The models have a constant effective population size of  $N_0 = 10,000$  up to 1,000 generations ago, at which time the population expands to size  $N$  and remains at this size until the present. For the global model (GI), the effective population size is 10,000 until 5,000 generations before present, and then expands exponentially to a present size of  $5 \times 10^9$ . Note that expansion has relatively little effect on both LD and nucleotide diversity.
- **LD and nucleotide diversity for demographic models considered by Reich et al.<sup>21</sup>** (Panel c.) Models have a constant effective population size of 10,000 until 5,000 generations ago, at which time the population expands to  $10^8$ . The population then undergoes a contraction to an effective size of  $N$  at the time of  $800 + 0.4N$  generations (first three models) or  $800 + 0.8N$  generations before the present, and then expands once again to  $10^8$  at 800 generations before the present, and remains at that size until the present. Note that nucleotide diversity for all six models is much lower than the observed human value of  $\sim 8 \times 10^{-4}$ .
- **Nucleotide diversity is plotted against half-distance of  $D'$  for above models.** (Panel d.) The models are labelled as C (constant effective population size of 10,000), GI (global model), B1–5 (bottlenecks with contraction factor of 0.5 to 0.01, panel a), E1–5 (expansion by a factor of 2 to 100, panel b) and R1–6 (the six models, panel c). Note the close inverse correlation between the extent of LD and nucleotide diversity.



the recent studies have mostly focused on SNPs (see TABLE 1 for a summary of many recent studies). We will not review studies of microsatellites in detail here — SNPs are more likely to be useful for population association studies as a result of their much higher density in

the genome and much greater mutational stability. Additionally, because microsatellites are multiallelic, a satisfactory measure of LD has not been derived. Most studies have used the *p*-values of a Fisher's exact test as a proxy for LD, but it is important to note that this is a test

Table 1 | **Some recent empirical studies of linkage disequilibrium in humans**

Marker type	Population(s)	Sample number (people)	Number of markers	Region type	Number of regions	Total size of region surveyed (kb)	Chromosome(s)	Refs
SNP	African American, Finnish and non-Hispanic white	71	88	Gene: <i>LPL</i>	1	9.7	8	15
SNP	European American and African American	11	78	Gene: <i>ACE</i>	1	24	17	58
SNP	CEPH Four ethnic groups	89 (9 families) 260	14	Gene: <i>ATM</i>	1	142	11	56
SNP	European CEPH Finnish Sardinian	92 males 100 males 150 males	39 (more within each region)	Random genomic	2	1.34 Mb	Xq25 and Xq28	57
SNP micro	British	810 (158 families)	24 SNPs 2 micros	Variable segments <i>TCRA</i>	1	850	14	17
SNP	Afrikaner, Ashkenazim, Finnish and British	~1,600	38	Genes and random genomic	3	1.8 Mb	13, 19, 22	16
SNP	Global: 29 populations	~50 individuals each	4	Gene: <i>PAH</i>	1	75	12	54
SNP	Global	47	325	Random genomic	103	22	Multiple	18
SNPs VNTR micro	British	800 (160 families)	127	Genes	3	2.6 Mb	2, 13, 14	20
SNP/STRP	Global: 18 populations	972	2 SNP 4 STRP	Gene: <i>CFTR</i>	1	163	7	60
SNP	African American, Chinese, Mexican and North European	10	279	Variable segment <i>TCRB</i>	1	553	7	62
SNP	Utah-CEPH Yoruban (Nigerian)	44 96	272	Random genomic	19	251	Multiple	21
SNP	Caucasian, African American, Asian, Hispanic and Native American	82	3,899	Gene <sup>†</sup>	313	720	Multiple	22
SNP	Unrelated North European men	50	179	HLA	1	216	6	66
SNP	European	384	122	9 genes	–	135	Multiple	67
SNP	European	129 trios	103	Region that spans several genes	1	500	5	65
SNP	African, Asian and Caucasian	20 haploid chromosomes (~10 individuals)	~24,000	Whole-chromosome non-repetitive DNA	–	~21.7 Mb	21 (entire chromosome)	52
SNP	Japanese Utah-CEPH	77	44	Gene: <i>AGT</i>	1	14.4	1	61

\*Linkage disequilibrium examined at distinct subregions around a core single-nucleotide polymorphism (SNP) within the coding region of a gene — regions each ~2 kb, spaced at 0, 5, 10, 20, 40, 80 and 160 kb from the core SNP. †Gene represents the transcriptional unit of each gene including coding region, 5' UTR, 3' UTR, 100 bp of intron–exon boundary, 5' upstream (not a fully contiguous region). *ACE*, angiotensin-I-converting enzyme (peptidyl-dipeptidase A) 1; *AGT*, angiotensinogen (serine (or cysteine) proteinase inhibitor, clade A ( $\alpha$ -1 antiproteinase, antitrypsin), member 8); *ATM*, ataxia telangiectasia mutated; CEPH, families from the collection at the Centre d'Etude du Polymorphisme Humaine; *CFTR*, cystic fibrosis transmembrane conductance regulator, ATP-binding cassette (sub-family C, member 7); HLA, human leukocyte antigen; *LPL*, lipoprotein lipase; micro, microsatellite; *PAH*, phenylalanine hydroxylase; STRP, simple tandem repeat polymorphism; *TCRA*, T-cell receptor- $\alpha$ ; *TCRB*, T-cell receptor- $\beta$ ; VNTR, variable number of tandem repeats.



statistic, not a measure of LD (see above); the adequacy of this statistic as a quantitative measure of LD has not been convincingly shown. With their greater genotyping ease, microsatellites have provided the possibility of surveying entire genomes, albeit at fairly low density<sup>51</sup>. Interestingly, these large-scale surveys seem to find allelic associations (LD) between occasional loci that are separated by many megabases, although a consistent dependence of LD on physical or genetic distance has not been very clear. The occasional finding of LD among microsatellites that are separated by such vast stretches is not consistent with current SNP surveys, and the explanation for this is unclear, although it might involve a greater power to detect LD between multi-allelic markers as well as differences in measures of LD<sup>24</sup>.

In most surveys of LD among SNPs, the populations studied have typically been European, or of European descent, although globally diverse collections of individuals have been studied, as have populations thought to have been isolated and derived from a small number of founders (for example, Sardinia, Finland and Iceland) (see TABLE 1). Sample sizes vary from as few as ten to well over 1,000 individuals, and, like the *Drosophila* studies, the distances over which LD has been examined have varied from focal gene-based studies to regions that stretch over several megabases, and in the most recent instance, to an entire chromosome<sup>52</sup> (see TABLE 1). Similar to the *Drosophila* findings, considerable variation in the pattern and extent of LD has been found.

Reviews of published data show that LD varies among populations, with European populations typically showing lower nucleotide diversity and greater LD than African populations<sup>21,28,53,54</sup>. The strength of LD and the distance over which it extends also vary from one region of the genome to another. Comparisons are complicated by the fact that different studies have used different measures of LD, as well as different values of those measures for defining LD as 'useful', 'present' or 'strong' (see above). There are reports of LD between markers that are separated by distances of >100 kb (REFS 17,19–22,51,55,56), with some claims of extensive LD across regions of up to 1 Mb (REF. 57). By contrast, an equivalent number of studies have found weak LD at much shorter distances<sup>15–18,54,58–61</sup>. Reich *et al.*<sup>21</sup>, studying 19 randomly selected genomic regions, reported that, on average, LD in a population of Utah Mormons, generally of north-European descent, has an average half-length of about 60 kb around very common alleles (>0.35 allele frequency) when measured by *D'*. This corresponds to an extent of 'useful LD' of ~20 kb (see above). Much weaker LD was found in an African population. The authors proposed a strong population bottleneck during the history of northern Europeans as an explanation for the greater extent of LD, but such an explanation is inconsistent with the high observed nucleotide diversity of this population (BOX 2c). Methodological problems in study design and analysis might have affected the conclusions of this study, as recently discussed in a detailed critique by Weiss and Clark<sup>25</sup>. One hallmark of all these studies is that the variation in LD at all distances is great, and is not

predictable from one region to another<sup>20–22,62</sup>. Abecasis *et al.*<sup>20</sup> estimated that physical distance could account for less than 50% of the variation in LD in their study. They suggested that the remaining variation was probably due to a combination of drift, demographic factors, selection and variable rates of mutation, recombination and gene conversion. This will probably be the rule, rather than the exception, and the relative contributions of these forces will vary regionally, which indicates that great care will be needed in any prediction of LD in regions where it has not yet been empirically assessed.

#### Are there discrete 'blocks' of LD?

Patterns in the chromosomal organization of LD have become more apparent. Early studies had noted tracts of long homozygous segments<sup>63</sup>, or regions of strong LD juxtaposed between regions where markers were in equilibrium<sup>57</sup>. New data indicate that this might result from the structuring of LD into discrete blocks of haplotypes that show high LD, separated by possible hot spots of recombination and a breakdown of LD<sup>64–68</sup>. However, it has also been shown that block-like patterns of LD can result from stochastic fluctuations even when recombination rate is uniform across the genome<sup>62</sup>. At least one whole chromosome seems to show a block-like structure, with haplotype diversity being extremely limited in blocks that range in size from a few kilobases to, in some instances, >100 kb (REF. 52), although these conclusions are limited by a sample of only 20 chromosomes. Another study of haplotype structure in genes that were sequenced in a larger population found that, although most genes had a single haplotype with a frequency of  $\geq 50\%$ , one-third of the genes studied did not have a single predominant haplotype<sup>22</sup>. This again indicates that LD should be empirically assessed in a region before disease-mapping studies, rather than extrapolating from other regions, so that an appropriate density of markers can be used<sup>69</sup>.

#### Conclusions

The debate over the average extent of levels of LD that are useful for association mapping is becoming narrower as data become available for more genomic regions and populations. Although even the authors of this review are not in total agreement, it seems safe to propose a range of 10–30 kb for the extensively studied northern European populations. The extent seems to be less in several African populations<sup>21,28</sup>, and remains little studied for other parts of the world. Perhaps more importantly, it has become clear that the average extent might not be a good guide for the design and feasibility of LD mapping approaches. This is true for at least two reasons. First, the tremendous variability in the extent of LD from one region of the genome to another means that the average will greatly overstate the useful range for some regions and understate it for others. Second, even in a region of high mean LD, some pairs of loci do not show useful levels of LD due to gene conversion, differences in allele frequency and perhaps other factors. An important and almost entirely unanswered question is whether the patterns of LD found in one population

will be replicated in other populations with differing population histories. What little data that can be applied to answering this question are conflicting, with hints that patterns of LD are similar among different populations equalled by indications that each population is substantially different. Answering this question, and establishing the generality (or not) of haplotype maps constructed in one population, should be an urgent priority for research. It is also worth briefly noting that the use of LD for mapping relies on assumptions regarding the genetic architecture of common diseases that are open to question; this point is discussed extensively elsewhere<sup>70–72</sup>.

A less practical but perhaps more interesting question is what forces have shaped the patterns of LD in humans. An increasingly persuasive case can be made that simple demographic models of population expansions and contractions are insufficient to explain the observed patterns. More complex historical models might do better, but molecular forces, such as gene conversion<sup>18,28,73,74</sup> and recombination hot spots<sup>66</sup>, have also recently come to the fore. Selection — positive, negative or balancing — must also have had an influence, but its role has been difficult to show conclusively. Sorting out these factors might occupy students of LD long after its more utilitarian uses have played themselves out.

1. Boehnke, M. Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *Am. J. Hum. Genet.* **55**, 379–390 (1994).
2. Kruglyak, L. & Lander, E. S. High-resolution genetic mapping of complex traits. *Am. J. Hum. Genet.* **56**, 1212–1223 (1995).
3. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
4. Hastbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
5. Jorde, L. B. Linkage disequilibrium as a gene mapping tool. *Am. J. Hum. Genet.* **56**, 11–14 (1995).
6. Jorde, L. B. *et al.* Linkage disequilibrium in the neurofibromatosis 1 (*NF1*) region: implications for gene mapping. *Am. J. Hum. Genet.* **53**, 1038–1050 (1993).
7. Jorde, L. B. *et al.* Linkage disequilibrium predicts distance in the adenomatous polyposis coli region. *Am. J. Hum. Genet.* **54**, 884–898 (1994).
8. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
9. Watkins, W. S. Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. *Am. J. Hum. Genet.* **55**, 348–355 (1994).
10. Sachidanandam, R. *et al.* A map of the human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
11. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
12. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
13. Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet.* **2**, 930–942 (2001).
14. Lewontin, R. C. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49–67 (1964).
15. Clark, A. G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
16. Dunning, A. M. *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67**, 1544–1554 (2000).
17. Moffatt, M. F., Trahern, J. A., Abecasis, G. R. & Cookson, W. O. Single nucleotide polymorphism and linkage disequilibrium within the *TCR $\alpha$ / $\delta$*  locus. *Hum. Mol. Genet.* **9**, 1011–1019 (2000).
18. Ardlie, K. *et al.* Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.* **69**, 582–589 (2001).
19. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
20. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
21. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
22. Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
23. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
24. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).  
**A comprehensive review that compares empirical data with population models, from a population genetic perspective. Discusses different measures of LD.**
25. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).  
**An excellent critical survey of the issues that surround LD mapping and SNP map design.**
26. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
27. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
28. Frisse, L. *et al.* Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**, 831–843 (2001).
29. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
30. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401 (2000).
31. Long, A. D. & Langley C. H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **8**, 720–731 (1999).  
**A good introduction to the population genetic parameter  $4N_e c$  as a measurement by which to compare LD across different regions.**
32. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942 (2000).
33. Wall, J. D. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17**, 156–163 (2000).
34. Zapata, C. & Alvarez, C. On the detection of nonrandom associations between DNA polymorphisms in natural populations of *Drosophila*. *Mol. Biol. Evol.* **10**, 823–841 (1983).
35. Schaeffer, S. W. & Miller, E. L. Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **135**, 541–542 (1993).
36. Langley, C. H. *et al.* Linkage disequilibrium and the site frequency spectra in the *su(s)* and *su(w<sup>2</sup>)* regions of the *Drosophila melanogaster* X chromosome. *Genetics* **156**, 1837–1852 (2000).
37. Schaeffer, S. W. *et al.* Protein variation in ADH and ADH-RELATED in *Drosophila pseudoobscura*: linkage disequilibrium between single nucleotide polymorphisms and protein alleles. *Genetics* **159**, 673–687 (2001).
38. Wang, W., Thornton, K., Berry, A. & Long, M. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science* **295**, 134–137 (2002).
39. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
40. Begun, D. J. & Aquadro, C. F. Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of *Drosophila*: selection and geographic differentiation. *Genetics* **136**, 155–171 (1994).
41. Parsch, J., Meiklejohn, C. D. & Hartl, D. L. Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of *Drosophila simulans*. *Genetics* **159**, 647–657 (2001).
42. Verrelli, B. C. & Eanes, W. F. Clinal variation for amino acid polymorphisms at the *Pgm* locus in *Drosophila melanogaster*. *Genetics* **157**, 1649–1663 (2001).
43. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1983).
44. Andolfatto, P. & Przeworski, M. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156**, 257–268 (2000).
45. Payseur, B. A. & Nachman, M. W. Microsatellite variation and recombination rate in the human genome. *Genetics* **156**, 1285–1298 (2000).
46. Nachman, M. W. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**, 481–485 (2001).
47. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–237 (1999).
48. Wakeley, J., Nielsen, R., Liu-Cordero, S. N. & Ardlie, K. The discovery of single nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet.* **69**, 1332–1347 (2001).
49. Peterson, A. C. *et al.* The distribution of linkage disequilibrium over anonymous genome regions. *Hum. Mol. Genet.* **4**, 887–894 (1995).
50. Laan, M. & Paabo, S. Demographic history and linkage disequilibrium in human populations. *Nature Genet.* **4**, 435–438 (1997).
51. Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722 (1999).
52. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* **294**, 1719–1722 (2001).  
**Describes an innovative and exciting experimental approach to whole-chromosome/genome studies of sequence variation and LD, and finds surprisingly limited haplotype diversity over much of chromosome 21 in a global sample.**
53. Goddard, K. A., Hopkins, P. J., Hall, J. M. & Witte, J. S. Linkage disequilibrium and allele frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216–234 (2000).
54. Kidd, J. R. *et al.* Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH*, in a global representation of populations. *Am. J. Hum. Genet.* **66**, 1882–1899 (2000).
55. Lonjou, C., Collins, A. & Morton, N. E. Allelic association between marker loci. *Proc. Natl Acad. Sci. USA* **96**, 1621–1626 (1999).
56. Bonnen, P. E. *et al.* Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. *Am. J. Hum. Genet.* **67**, 1437–1451 (2000).
57. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
58. Rieder, M. J., Taylor, S. L., Clark, A. G. & Nickerson, D. A. Sequence variation in the human angiotensin converting enzyme. *Nature Genet.* **22**, 59–62 (1999).
59. Templeton, A. R. *et al.* Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics* **156**, 1259–1275 (2000).

60. Mateu, E. *et al.* Worldwide genetic analysis of the *CFTR* region. *Am. J. Hum. Genet.* **68**, 103–117 (2001).
61. Nakajima, T. *et al.* Nucleotide diversity and haplotype structure of the human angiotensinogen gene in two populations. *Am. J. Hum. Genet.* **70**, 109–123 (2002).
62. Subrahmanyam, L. *et al.* Sequence variation and linkage disequilibrium in the human T-cell receptor  $\beta$  (*TCRB*) locus. *Am. J. Hum. Genet.* **69**, 381–395 (2001).
63. Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the center d'Étude du Polymorphisme Humain. *Am. J. Hum. Genet.* **65**, 1493–1500 (1999).
64. Goldstein, D. B. Islands of linkage disequilibrium. *Nature Genet.* **29**, 109–111 (2001).
65. Daly, M. J. *et al.* High resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
66. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
67. Johnson, G. C. L. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
68. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohns disease. *Nature Genet.* **29**, 223–228 (2001).
69. Martin, E. R. *et al.* SNPping away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* **67**, 383–394 (2000).
70. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
71. Weiss, K. M. & Terwilliger, J. D. How many SNPs does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157 (2000).
- A clear and critical discussion of the very real difficulties that are inherent in identifying genes that contribute to complex traits.**
72. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
73. Przeworski, M. & Wall, J. D. Why is there so little intragenic linkage disequilibrium in humans? *Genet. Res.* **77**, 143–151 (2001).
74. Quintana, P. J. E., Neuwirth, E. A. H. & Grosovsky, A. J. Interchromosomal gene conversion at an endogenous human cell locus. *Genetics* **158**, 757–767 (2001).
75. Terwilliger, J. D., Zollner, S., Laan, M. & Paabo, S. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum. Hered.* **48**, 138–154 (1998).
76. Chakraborty, R. & Weiss, K. M. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl Acad. Sci. USA* **85**, 9119–9123 (1988).
77. Stephens, J. C., Briscoe, D. & O'Brien, S. J. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am. J. Hum. Genet.* **55**, 809–824 (1994).
78. Wilson, J. F. & Goldstein, D. B. Consistent long-range linkage disequilibrium generated by admixture in a Bantu–Semitic hybrid population. *Am. J. Hum. Genet.* **67**, 926–935 (2000).
79. Pfaff, C. L. *et al.* Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.* **68**, 198–207 (2001).
80. Nordborg, M. *et al.* The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genet.* **30**, 190–193 (2002).
81. Cannon, G. B. The effects of natural selection on linkage disequilibrium and relative fitness in experimental populations of *Drosophila melanogaster*. *Genetics* **48**, 1201–1216 (1963).

#### Acknowledgements

We are grateful to M. Eberle for carrying out the simulations presented in Box 2. The manuscript was greatly improved by suggestions from three anonymous referees. Supported in part by a grant from the National Institutes of Mental Health (to L.K.). L.K. is a James S. McDonnell Centennial Fellow. M.S. is supported by the National Human Genome Research Institute.

#### Online links

##### DATABASES

The following terms in this article are linked online to:

LocusLink: <http://www.ncbi.nlm.nih.gov/LocusLink/Pgm>

##### FURTHER INFORMATION

SNP Consortium: <http://snp.cshl.org>

Access to this interactive links box is free online.